Projections, Radon-Nikodym, and conditioning

1 Projections in $\mathcal{L}^2$ 1
2 Radon-Nikodym theorem 3
3 Conditioning 5
3.1 Projections as conditional expectation maps 6
4 Bringing it all back to Ω 8

1 Projections in $\mathcal{L}^2$

For a measure space $(\mathcal{X}, \mathcal{A}, \mu)$, the set $\mathcal{L}^2 = \mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mu)$ of all square integrable, $\mathcal{A}\backslash\mathcal{B}(\mathbb{R})$-measurable real functions on $\mathcal{X}$ would be a Hilbert space if we worked with equivalence classes of functions that differ only on μ-negligible sets. The corresponding set $L^2(\mathcal{X}, \mathcal{A}, \mu)$ can be identified as a true Hilbert space.

Lazy probabilists (like me) often ignore the distinction between L^2 and $\mathcal{L}^2$, referring to $\|f\|_2 = \big(\mu(f^2)\big)^{1/2}$ as a norm on $\mathcal{L}^2$ (rather than using the more precise term 'semi-norm') and

$$\langle f, g\rangle = \mu(fg) \qquad \text{for } f, g \in \mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mu)$$

as an inner product. It is true that $\langle f, g\rangle$ is linear in f for fixed g and linear in g for fixed f, and it is true that $\|f\|^2 = \langle f, f\rangle$, but we can only deduce that $f(x) = 0$ a.e.$[\mu]$ if $\|f\|_2 = 0$. As shown by HW3.2, the space $\mathcal{L}^2$ is also complete: for each Cauchy sequence $\{h_n : n \in \mathbb{N}\}$ in $\mathcal{L}^2$ there exists an h in $\mathcal{L}^2$ (unique up to μ-equivalence) for which $\|h_n - h\|_2 \to 0$.

To avoid some tedious qualifications I will slightly abuse terminology by referring to a subset $\mathcal{H}$ of $\mathcal{L}^2$ as ***closed*** if: for each f in $\mathcal{L}^2$ with $\|h_n - f\|_2 \to 0$ for a sequence $\{h_n\}$ in $\mathcal{H}$ there exists an h in $\mathcal{H}$ for which $h(x) = f(x)$ a.e.$[\mu]$. (Some authors would insist that f itself should belong to $\mathcal{H}$.)

The following result underlies the existence of both Radon-Nikodym derivatives (densities) for measures and Kolmogorov conditional expectations.

<*1*> **Theorem.** *Suppose $\mathcal{H}$ is a closed subspace of $\mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mu)$. For each $f \in \mathcal{L}^2$ there exists an f_0 in $\mathcal{H}$ for which:*

(i) $\|f - f_0\|_2 = \delta := \inf\{\|f - h\|_2 : h \in \mathcal{H}\}$*;*

(ii) $\langle f - f_0, h\rangle = 0$ *for every h in $\mathcal{H}$;*

(iii) property (iii) uniquely determines f_0 up to μ-equivalence;

(iv) $\|f\|_2^2 = \|f_0\|_2^2 + \|f - f_0\|_2^2$.

Proof The argument uses completeness of $\mathcal{L}^2$ and the identity

<2> $$\|a+b\|_2^2 + \|a-b\|_2^2 = 2\|a\|_2^2 + \|b\|_2^2 \qquad \text{for all } a, b \in \mathcal{L}^2,$$

an equality that results from expanding $\langle a+b, a+b\rangle + \langle a-b, a-b\rangle$ then cancelling out $\langle a, b\rangle$ terms.

By definition of the infimum, for each $n \in \mathbb{N}$ there exists an $h_n \in \mathcal{H}$ for which

$$\|f - h_n\|_2 \le \delta_n := \delta + n^{-1}.$$

Invoke equality <2> with $a = f - h_n$ and $b = f - h_m$:

$$4\|f - (h_n + h_m)/2\|_2^2 + \|h_n - h_m\|_2^2 = 2\|f - h_n\|_2^2 + 2\|f - h_m\|_2^2.$$

The first term on the left-hand side is $\ge 4\delta^2$ because $(h_n + h_m)/2 \in \mathcal{H}$. Thus

$$\|h_n - h_m\|_2^2 \le 2\delta_n^2 + 2\delta_m^2 - 4\delta^2 \to 0 \qquad \text{as } \min(m, n) \to \infty.$$

That is $\{h_n\}$ is a Cauchy sequence, which converges in norm to an f_0 in $\mathcal{L}^2$. Without loss of generality ($\mathcal{H}$ is closed) we may assume that $f_0 \in \mathcal{H}$.

Equality (i) follows from

$$\delta \le \|f - f_0\|_2 \le \|f - h_n\|_2 + \|h_n - f_0\|_2 \to \delta \qquad \text{as } n \to \infty.$$

For (ii) note, for each $h \in \mathcal{H}$, that the quadratic

$$\|f - (f_0 + th)\|^2 = \|f - f_0\|^2 + 2t\langle f - f_0, h\rangle + t^2 \|h\|_2^2$$

achieves its minimum value δ^2 at $t = 0$, which forces the coefficient of t to equal zero.

For (iii) suppose $f_0, f_1 \in \mathcal{H}$ and both $f - f_0$ and $f - f_1$ are orthogonal to each h in $\mathcal{H}$. Then the difference $f_0 - f_1$ must be orthogonal to itself, that is $\|f_0 - f_1\|_2^2 = 0$, forcing $f_0 = f_1$ a.e.$[\mu]$.

Equality (iv) follows from the fact that $\langle f - f_0, f_0\rangle = 0$.

□

The function f_0, which is unique only up to μ-equivalence, is called an (orthogonal) projection of f onto $\mathcal{H}$. Formally the projection function $\pi_{\mathcal{H}}$ maps an f from $\mathcal{L}^2(\mu)$ to a μ-equivalence classes of functions in $\mathcal{H}$. If we arbitrarily choose one member from each equivalence class then $\pi_{\mathcal{H}}$ can also be thought of as a map from $\mathcal{L}^2(\mu)$ into $\mathcal{H}$, at the cost of some caveats involving negligible sets. For example, if $g_1, g_2 \in \mathcal{L}^2$ and $\pi_{\mathcal{H}} g_i = h_i$ for $i = 1, 2$ then, for constants c_1 and c_2, part (iii) of the Theorem gives

<3> $$\pi_{\mathcal{H}}(c_1 g_1 + c_2 g_2) = c_1 \pi_{\mathcal{H}} g_1 + c_2 \pi_{\mathcal{H}} g_2 \quad \text{a.e.}[\mu],$$

which is as close to linearity as we can hope to get for a map that is only defined up to a μ-equivalence.

2 Radon-Nikodym theorem

The simplest form of the theorem concerns two finite measures μ and ν defined on some $(\mathcal{X}, \mathcal{A})$.

<4> **Theorem.** *If $\mu\mathcal{X} < \infty$ and $\mu f \ge \nu f$ for each f in $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$ then there exists an $\mathcal{A}$-measurable function Δ with $0 \le \Delta(x) \le 1$ for all x such that*

$$\nu f = \mu(f\Delta) \qquad \text{for each } f \text{ in } \mathcal{M}^+(\mathcal{X}, \mathcal{A}).$$

The function Δ is unique up to μ-equivalence.

Proof (sketch of a proof due to von Neumann, 1940, page 127)

Without loss of generality suppose $\nu\mathcal{X} = 1$.

Define $\mathcal{H} = \{f \in \mathcal{L}^2(\mathcal{X}, \mathcal{A}, \mu) : \nu f = 0\}$. Note that $\mathcal{H}$ is a closed subspace of $\mathcal{L}^2$: if $\nu h_n = 0$ and $\mu|h_n - f|^2 \to 0$ then

$$|\nu f| = |\nu(h_n - f)| \le \nu|h_n - f| \le \sqrt{\nu|h_n - f|^2} \le \sqrt{\mu|h_n - f|^2} \to 0,$$

implying $\nu f = 0$.

Let $f_0 = \pi_{\mathcal{H}} 1$ and $g = 1 - f_0$. Note that $\nu g = \nu 1 - \nu f_0 = 1$. In consequence $\mu\{x : g(x) \neq 0\} \geq \nu\{x : g(x) \neq 0\} > 0$ and $\|g\|_2^2 := \mu(g^2) > 0$. Suppose $f \in \mathcal{L}^2$ has $\nu f = c$. Then $\nu(f - cg) = 0$, that is, $f - cg \in \mathcal{H}$, so that $0 = \langle f - cg, g\rangle = \mu(fg - cg^2) = 0$. The final equality rearranges to

$$\nu f = \mu(f\Delta) \qquad \text{where } \Delta := g/\|g\|_2^2.$$

Invoke the last equality with $f = \{\Delta < 0\}$ to get

$$0 \leq \nu\{\Delta < 0\} \leq \mu\left(\Delta\{\Delta < 0\}\right) \leq 0,$$

with the last inequality strict unless $\mu\{\Delta < 0\} = 0$. Similarly

$$\mu\{\Delta > 1\} \geq \nu\{\Delta > 1\} = \mu\left(\Delta\{\Delta > 1\}\right) \geq \mu\{\{\Delta > 1\},$$

with the last inequality strict unless $\mu\{\Delta > 1\} = 0$.

Replace Δ by $\Delta\{0 \leq \Delta \leq 1\}$.

For $f \in \mathcal{M}^+(\mathcal{X}, \mathcal{A})$ take limits (MC) in $\nu(f \wedge n) = \mu\Delta(f \wedge n)$ as $n \to \infty$.

For uniqueness a.e.$[\mu]$: If $\mu(f\Delta_1) = \mu(f\Delta_2)$ for all $f \in \mathcal{M}^+$ consider first $f = \{\Delta_1 < \Delta_2\}$ then $f = \{\Delta_1 > \Delta_2\}$ to deduce that $\Delta_1 = \Delta_2$ a.e.$[\mu]$.

□

Theorem <4> has an extension to sigma-finite measures with ν ***dominated*** by μ, that is: for each $A \in \mathcal{A}$, if $\mu A = 0$ then $\nu A = 0$.

> **Remark.** Domination is sometimes expressed as "ν is absolutely continuous with respect to μ", which is often denoted by $\nu \ll \mu$. This terminology borrows from the classical concept of absolute continuity of a function defined on the real line (Pollard, 2001, Section 3.4).

<5> **Theorem.** *If μ and ν are both sigma-finite measures with ν dominated by μ then there exists a real-valued function $\Delta \in \mathcal{M}^+(\mathcal{X}, \mathcal{A})$ for which*

$$\nu f = \mu(f\Delta) \qquad \text{for each } f \text{ in } \mathcal{M}^+(\mathcal{X}, \mathcal{A}).$$

The function Δ is unique up to μ-equivalence.

For an idea of the proof see Pollard (2001, Section 3.2).

3 Conditioning

Recall the conditioning problem. We have a probability measure $\mathbb{Q}$ on the product sigma-field $\mathcal{A}\otimes\mathcal{B}$ on $\mathcal{X}\times\mathcal{Y}$, with $\mathcal{X}$-marginal P. That is,

$$Pg = \mathbb{Q}^{x,y}g(x) \qquad \text{for each } g\in\mathcal{M}^+(\mathcal{X},\mathcal{A}).$$

Here and subsequently I identify g with a function on $\mathcal{X}\times\mathcal{Y}$ whose value does not depend on the $\mathcal{Y}$-coordinate.

We seek a Markov kernel $\mathbb{K} = \{\mathbb{K}_x : x\in\mathcal{X}\}$—a family of probability measures on $\mathcal{B}$ for which $x\mapsto\mathbb{K}_xB$ is $\mathcal{A}$-measurable for each $B\in\mathcal{B}$—for which

<6> $$\mathbb{Q}f(x,y) = P^x\mathbb{K}_x^y f(x,y) \qquad \text{for each } f\in\mathcal{M}^+(\mathcal{X}\times\mathcal{Y},\mathcal{A}\otimes\mathcal{B}).$$

We can also think of $\mathbb{K}$ as a map from $\mathcal{M}^+(\mathcal{X}\times\mathcal{Y},\mathcal{A}\otimes\mathcal{B})$ into $\mathcal{M}^+(\mathcal{X},\mathcal{A})$ by taking $\mathbb{K}f$ to be the function whose value at x equals $\mathbb{K}_xf = \mathbb{K}_x^y f(x,y)$. We require this map to have the following ***Markov kernel properties***:

(i) $\mathbb{K}_x0 = 0$ and $\mathbb{K}_x1 = 1$;

(ii) $\mathbb{K}_x(c_1f_1 + c_2f_2) = c_1\mathbb{K}_xf_1 + c_2\mathbb{K}_xf_2$ for constants $c_i\in\mathbb{R}^+$;

(iii) $\mathbb{K}_xf_1 \le \mathbb{K}_xf_2$ if $f_1(x,y)\le f_2(x,y)$ for all (x,y);

(iv) if $f_n(x,y)\uparrow f(x,y)$ then $\mathbb{K}_x^yf_n(x,y)\to\mathbb{K}_x^yf(x,y)$.

(v) if $g\in\mathcal{M}^+(\mathcal{X},\mathcal{A})$ and $f\in\mathcal{M}^+(\mathcal{X}\times\mathcal{Y},\mathcal{A}\otimes\mathcal{B})$ then $\mathbb{K}_x^y\left(g(x)f(x,y)\right) = g(x)\mathbb{K}_xf$.

Property (v) is more a statement of the fact that $\mathbb{K}_x$ treats $g(x)$ like a constant than a requirement that needs to be checked.

As you will soon see, projections can be used to define a ***conditional expectation map*** $\mathcal{K} : \mathcal{M}^+(\mathcal{X}\times\mathcal{Y},\mathcal{A}\otimes\mathcal{B}) \to \mathcal{M}^+(\mathcal{X},\mathcal{A})$ with analogous properties:

(i) $\mathcal{K}_x0 = 0$ and $\mathcal{K}_x1 = 1$ a.e.$[P]$;

(ii) $\mathcal{K}_x(c_1f_1 + c_2f_2) = c_1\mathcal{K}_xf_1 + c_2\mathcal{K}_xf_2$ a.e.$[P]$ for constants $c_i\in\mathbb{R}^+$;

(iii) $\mathcal{K}_xf_1\le\mathcal{K}_xf_2$ a.e.$[P]$ if $f_1(x,y)\le f_2(x,y)$ for all (x,y);

(iv) if $f_n(x,y)\uparrow f(x,y)$ then $\mathcal{K}_x^yf_n(x,y)\uparrow\mathcal{K}_x^yf(x,y)$ a.e.$[P]$;

(v) if $g \in \mathcal{M}^+(\mathcal{X}, \mathcal{A})$ and $f \in \mathcal{M}^+(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$ then $\mathcal{K}_x^y(gf) = g(x)\mathcal{K}_x f$ a.e.$[P]$.

If the a.e.$[P]$ were not added to each line, $\mathcal{K}$ would correspond to a Markov kernel. If we collect the (uncountably many) P-negligible sets into a single P-negligible set $\mathcal{N}$ then a redefinition of $\mathcal{K}_x$ for $x \in \mathcal{N}$ could provide a Markov kernel. Unfortunately reduction to a single P-negligible $\mathcal{N}$ is not always possible and even when it is possible it takes a lot of work. Alternatively we could just learn to live with all the a.e.$[P]$ constraints and accept something less than a Markov kernel for the purposes of conditioning. That is the choice made in much of the probability and statistics literature, with the $\mathcal{K}$ sometimes being referred to as a ***Kolmogorov conditional expectation*** operator.

3.1 Projections as conditional expectation maps

Here is how projections get into the story. We can identify $\mathcal{L}^2(P) := \mathcal{L}^2(\mathcal{X}, \mathcal{A}, P)$ with a subspace of $\mathcal{L}^2(\mathbb{Q}) := \mathcal{L}^2(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B}, \mathbb{Q})$ because $Pg(x)^2 = \mathbb{Q}g(x)^2$. In fact $\mathcal{L}^2(P)$ is then a closed subspace of $\mathcal{L}^2(\mathbb{Q})$. Indeed, suppose $\{g_n\}$ is a sequence in $\mathcal{L}^2(P)$ for which $\mathbb{Q}|g_n(x) - f(x,y)|^2 \to 0$ for some $f \in \mathcal{L}^2(\mathbb{Q})$. By an argument similar to HW2.2 and HW3.2, there exists a subsequence along which $g_{n(k)}(x) \to f(x,y)$ a.e.$[\mathbb{Q}]$. Deduce that f is $\mathbb{Q}$-equivalent to $\limsup_{n(k)} g_{n(k)}(x)$. Minor surgery on some P-negligible sets excludes values where the limsup equals $\pm\infty$, leaving a function in $\mathcal{L}^2(P)$.

If equality <6> holds, with $\mathbb{K}$ a Markov kernel, then $g(x) = \mathbb{K}_x^y f(x,y)$ is an $\mathcal{A}$-measurable function for which, by Jensen's inequality,

$$Pg^2 \le P(\mathbb{K}_x^y f(x,y)^2) \le \mathbb{Q}f^2 < \infty.$$

Moreover, for each $h \in \mathcal{L}^2(P)$,

$$\mathbb{Q}\left(f(x,y) - g(x)\right)h(x) = P^x\left(h(x)\mathbb{K}_x^y(f(x,y) - g(x)\right) = 0.$$

That is, $f - g$ is orthogonal to $\mathcal{H}$ in the sense of the $\mathcal{L}^2(\mathbb{Q})$ inner product, the property that identifies $g(x)$ as one of the $\mathcal{L}^2(P)$ functions that represents the orthogonal projection of f onto $\mathcal{L}^2(P)$.

Now try to go in the other direction. Let $\mathcal{K}$ denote the map that projects $\mathcal{L}^2(\mathbb{Q})$ orthogonally onto $\mathcal{L}^2(P)$, with $\mathcal{K}_x f$ denoting some arbitrarily choice from the P-equivalence class of possible functions. Of course, if $f(x,y) = g(x) \in \mathcal{L}^2(P)$ we can take $\mathcal{K}_x f$ to equal $g(x)$.

At the moment $\mathcal{K}$ is defined only on $\mathcal{L}^2(\mathbb{Q})$. A limiting operation will extend the definition to $\mathcal{M}^+(\mathcal{A} \otimes \mathcal{B})$. Let me first record how far we have gone towards defining a conditional expectation operator.

<7> **Lemma.** *The map $\mathcal{K}$ from $\mathcal{L}^2(\mathbb{Q})$ to $\mathcal{L}^2(P)$ has the following properties.*

(i) $\mathcal{K}_x 0 = 0$ *and* $\mathcal{K}_x 1 = 1$ a.e.$[P]$;

(ii) $\mathcal{K}_x(c_1 f_1 + c_2 f_2) = c_1 \mathcal{K}_x f_1 + c_2 \mathcal{K}_x f_2$ a.e.$[P]$ *for constants* $c_i \in \mathbb{R}^+$;

(iii) $\mathcal{K}_x f_1 \le \mathcal{K}_x f_2$ a.e.$[P]$ *if* $f_1(x,y) \le f_2(x,y)$ *for all* (x,y);

(iv) if $f_n(x,y) \uparrow f(x,y) \in \mathcal{L}^2(\mathbb{Q})$ *then* $\mathcal{K}_x^y f_n(x,y) \uparrow \mathcal{K}_x^y f(x,y)$ a.e.$[P]$

(v) $\mathbb{Q} g(x) f(x,y) = P g(x) \mathcal{K}_x f$ *for each* $g \in \mathcal{L}^2(P)$ *and* $f \in \mathcal{L}^2(\mathbb{Q})$.

Proof Assertion (i) holds even without the a.e.$[P]$ because constant functions belong to $\mathcal{L}^2(P)$.

Assertion (ii) corresponds to equality <3>.

For (iii) it suffices to prove that $g(x) := \mathcal{K}_x f \ge 0$ a.e.$[P]$ if $f(x,y) \ge 0$ for all (x,y). Use the fact that $f - g$ is orthogonal to $\mathbb{1}\{x : g(x) < 0\}$ because the indicator function belongs to $\mathcal{L}^2(P)$. Thus

$$\mathbb{Q} g\{g < 0\} = \mathbb{Q} f\{g < 0\} \ge 0.$$

The first integral would be strictly negative if $\mathbb{Q}\{g < 0\}$ were nonzero.

For (iv) note that $f - f_1 \ge f - f_n \downarrow 0$, so that $\mathbb{Q}(f - f_n)^2 \to 0$, by Dominated Convergence. Define $g(x) = \mathcal{K}_x f$ and $g_n(x) = \mathcal{K}_x f_n$. By (ii),

$$g_n(x) \uparrow G(x) := \sup\nolimits_{n \in \mathbb{N}} g_n(x) \le g(x) \quad \text{a.e.}[P].$$

A similar Dominated Convergence argument with $g - g_1 \ge G - g_n \downarrow 0$ then shows that $P|g_n(x) - G(x)|^2 \to 0$. Property (iv) from Theorem <1> shows that

$$P|g_n(x) - g(x)|^2 \le \mathbb{Q}|f - f_n|^2 \to 0.$$

It follows that $G(x) = g(x)$ a.e.$[P]$.

For (v) use the fact that $f - \mathcal{K}_x f$ is orthogonal to every function in $\mathcal{L}^2(P)$.

□

The extension to $\mathcal{M}^+(\mathcal{A} \otimes \mathcal{B})$ is now straightforward. To avoid notational confusion I will temporarily write $\widetilde{\mathcal{K}}$ for the extension.

<8> **Lemma.** *For* $f \in \mathcal{M}^+(\mathcal{A} \otimes \mathcal{B})$ *define* $\widetilde{\mathcal{K}}_x f = \sup_{i \in \mathbb{N}} \mathcal{K}_x(f \wedge i)$. *If* $f_n \in \mathcal{M}^+(\mathcal{A} \otimes \mathcal{B})$ *and* $f_n \uparrow f$ *then* $\widetilde{\mathcal{K}}_x f_n \uparrow \widetilde{\mathcal{K}}_x f$ a.e.$[P]$.

Proof By construction we have $\widetilde{\mathcal{K}}_x f_n \le \widetilde{\mathcal{K}}_x f_{n+1} \le \widetilde{\mathcal{K}}_x f$ a.e.$[P]$ for each n. Lemma <7> (iv) gives $\mathcal{K}_x(f_n \wedge i) \uparrow \mathcal{K}_x(f \wedge i)$ a.e.$[P]$, for each $i \in \mathbb{N}$. An interchange in the order of two suprema then gives

$$\sup_n \widetilde{\mathcal{K}}_x f_n = \sup_n \sup_i \mathcal{K}_x(f_n \wedge i) = \sup_i \mathcal{K}_x(f \wedge i) = \widetilde{\mathcal{K}}_x f \quad \text{a.e.}[P],$$

which is equivalent to the asserted convergence.

□

I leave it to you (HW8) to figure out how to combine Lemmas <7> and <8> to deduce that $\widetilde{\mathcal{K}}$ has all the desired properties for a Kolmogorov conditional expectation operator. It is also traditional to extend this operator to give a map from $\mathcal{L}^1(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B}, \mathbb{Q})$ to $\mathcal{L}^1(\mathcal{X}, \mathcal{A}, P)$, in much the same way that I extended integrals from $\mathcal{M}^+$ to $\mathcal{L}^1$.

4 Bringing it all back to Ω

The probability measure $\mathbb{Q}$ on $\mathcal{A} \otimes \mathcal{B}$ could be generated as the joint distribution of a real-valued random variable Y on $(\Omega, \mathcal{F}, \mathbb{P})$ and some arbitrary $\mathcal{F}\backslash\mathcal{A}$-measurable map X into $\mathcal{X}$.

Suppose $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ and $g \in \mathcal{L}^2(\mathcal{X}, \mathcal{A}, P)$. Then $f(x, y) = y$ belongs to $\mathcal{L}^2(\mathbb{Q})$ and $F(x) := \mathcal{K}_x f \in \mathcal{L}^2(P)$ and

$$\begin{aligned} \mathbb{P} g(X) Y &= \mathbb{Q}^{x,y} g(x) y \\ &= P^x g(x) \mathcal{K}_x y \\ &= P^x g(x) F(x) \\ &= \mathbb{P} g(X) F(X). \end{aligned}$$

If you can remember back to the third lecture of the course you should agree that if $\mathcal{G} = \sigma(X)$, the smallest sigma-field on Ω for which X is $\mathcal{G}\backslash\mathcal{A}$-measurable, then every function in $\mathcal{M}^+(\Omega, \mathcal{G})$ can be written in the form $h(X)$ for some h in $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$. In consequence, every random variable in $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ can be represented as $h(X)$ for some h in $\mathcal{L}^2(\mathcal{X}, \mathcal{A}, P)$. In particular, $g(X)$ and $F(X(\omega)) = \mathcal{K}_{X(\omega)} y$ from the last display both belong to $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ and the display is asserting that $Y - F(X)$ is orthogonal to $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$. That is, $F(X)$ is the orthogonal projection of Y onto $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$, which is actually a closed subspace of $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$.

In an obvious variation on traditional notation I write $\mathbb{P}_{\mathcal{G}}$ for the map (defined only up an almost sure equivalence) for which $F = \mathbb{P}_{\mathcal{G}} Y$. That is, $\mathbb{P}_{\mathcal{G}}$ projects $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ onto $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$. It is called the (Kolmogorov) conditional expectation of Y given the sub-sigma-field $\mathcal{G}$.

This operator can be extended to a map from $\mathcal{M}^+(\Omega, \mathcal{F})$ into (equivalence classes of) $\mathcal{M}^+(\Omega, \mathcal{G})$ with the properties

(i) $\mathbb{P}_{\mathcal{G}}0 = 0$ and $\mathbb{P}_{\mathcal{G}}1 = 1$ a.e.[$\mathbb{P}$];

(ii) $\mathbb{P}_{\mathcal{G}}(c_1 Y_1 + c_2 Y_2) = c_1 \mathbb{P}_{\mathcal{G}} Y_1 + c_2 \mathbb{P}_{\mathcal{G}} Y_2$ a.e.[$\mathbb{P}$] for constants $c_i \in \mathbb{R}^+$;

(iii) $\mathbb{P}_{\mathcal{G}} Y_1 \leq \mathbb{P}_{\mathcal{G}} Y_2$ a.e.[$\mathbb{P}$] if $Y_1(\omega) \leq Y_2(\omega)$ for all ω;

(iv) if $Y_n(\omega) \uparrow Y(\omega)$ then $\mathbb{P}_{\mathcal{G}} Y_n \uparrow \mathbb{P}_{\mathcal{G}} Y$ a.e.[$\mathbb{P}$];

(v) if $G \in \mathcal{M}^+(\omega, \mathcal{G})$ and $Y \in \mathcal{M}^+(\omega, \mathcal{F})$ then $\mathbb{P}_{\mathcal{G}}(GY) = G\mathbb{P}_{\mathcal{G}} Y$ a.e.[$\mathbb{P}$];

(vi) $\mathbb{P}Y = \mathbb{P}(\mathbb{P}_{\mathcal{G}} Y)$ a.e.[$\mathbb{P}$] for each $Y \in \mathcal{M}^+(\omega, \mathcal{F})$.

Again it is traditional to extend the projection $\mathbb{P}_{\mathcal{G}}$ to a map from $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ to $\mathcal{L}^1(\Omega, \mathcal{G}, \mathbb{P})$, defined only up to an almost sure equivalence, and having properties analogous to those of an integral.

You have probably noticed that X has disappeared from the notation, with only the fact that $\mathcal{G}$ equals $\sigma(X)$ left as a reminder. In fact the whole theory could be worked out for a completely general sub-sigma-field $\mathcal{G}$ of $\mathcal{F}$, with no mention of any X. For the special case when $\mathcal{G} = \sigma(X)$ the operator could also be written as $\mathbb{P}_X$.

Remark. If you want to be very cunning you could take $\mathcal{X} = \Omega$ with $\mathcal{A} = \mathcal{G}$, and X as the identity map, $X(\omega) = \omega$.

References

Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability.* Cambridge University Press.

von Neumann, J. (1940). On rings of operators. III. *Annals of Mathematics 41*(1), 94–161.